User Access & Information Protection

Managing Open Access & Information Protection

AIIM Industry White Paper on Records, Document and Enterprise Content Management for the Public Sector

First Edition 2002
ISBN 3-936534-00-4 (Industry White Papers Series)
ISBN 3-936534-04-7 (Industry White Paper 4)

Price (excl. VAT): 10 €

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Printed in United Kingdom by Stephens & George Print Group

User Access & Information Protection

Managing Open Access & Information Protection

AIIM Industry White Paper on Records, Document and Enterprise Content Management for the Public Sector

AIIM International Europe
Chappell House
The Green, Datchet
Berkshire SL3 9EH - UK
Tel: +44 (0)1753 592 769
Fax: +44 (0)1753 592 770
europeinfo@aiim.org

DLM-Forum
Electronic Records
Scientific Committee Secretariat
European Commission SG.B.3
Office JECL 3/36, Rue de la Loi 200, B-1049 Brussels - Belgium
Tel. +32 (0)2 299 59 00 / +32 (0) 2 295 67 21 / +32 (0)2 295 50 57
Fax +32 (0)2 296 10 95
A/e: dlm-forum@cec.eu.int

Author
IBM
Kim Jasper
Nymoellevej 91
DK-2800 Lyngby - Denmark
Tel. +45 (0)4586 5471
jasper@dk.ibm.com

Executive editors and coordinators
Dr. Ulrich Kampffmeyer
Silvia Kunze-Kirschner
PCI PROJECT CONSULT International Ltd.
Knyvett House, The Causeway
Staines, Middlesex TW18 3BA - UK
Tel.: +44 (0)1784 895 032
info@project-consult.com

Published by PROJECT CONSULT, Hamburg, 2002

Industry White Papers on Records, Document and Enterprise Content Management	Series	ISBN 3-936534-00-4
(1) Capture, Indexing & Auto-Categorization		ISBN 3-936534-01-2
(2) Conversion & Document Formats	HP	ISBN 3-936534-02-0
(3) Content Management	FileNET	ISBN 3-936534-03-9
(4) Access & Protection	IBM	ISBN 3-936534-04-7
(5) Availability & Preservation	Kodak	ISBN 3-936534-05-5
(6) Education, Training & Operation	TRW/ UCL/ comunicando	ISBN 3-936534-07-1

Preface

The Information Society impacts in many different ways on the European citizen, the most visible being the provision of access to information services and applications using new digital technologies. Economic competitiveness of Europe's technology companies and the creation of new knowledge-rich job opportunities are key to the emergence of a true European digital economy. Equally, the Information Society must reinforce the core values of Europe's social and cultural heritage – supporting equality of access, social inclusion and cultural diversity. One important element in ensuring a sound balance between these economic and social imperatives is co-operation between the information and communication industries and public institutions and administrations. Over the past 5 years, the European Commission in co-operation with EU Member States, has worked to create a multi-disciplinary platform for co-operation between technology providers and public institutions and administrations. The Forum aims at to make public administration more transparent, to better inform the citizen and to retain the collective memory of the Information Society. These objectives are at the heart of the eEurope Action Plan adopted by the European Summit in Feira on June 2000. I welcome the way the DLM-Forum has evolved over this period as a platform for identifying and promotion concrete solutions to many of the problems facing our public administrations.

In 1996 the initial focus of the DLM-Forum was on the guidelines for best practices for using electronic information and on dealing with machine-readable data and electronic documentation. More recently, at the last DLM-Forum in Brussels in 1999 a challenge was made to the ICT industries to assist public administrations in the EU Member States by providing proven and practical solutions in the field of electronic document and content management.

The importance of providing public access and long term preservation of electronic information is seen as a crucial requirement to preserve the "Memory of the Information Society" as well as improving business processes for more effective government. Solutions need to be developed that are, on the one hand, capable of adapting to rapid technological advances, while on the other hand guaranteeing both short and long term accessibility and the intelligent retrieval of the knowledge stored in document management and archival systems. Furthermore, training and educational programmes on understanding the technologies and standards used, as well as the identification of best practice examples, need to be addressed. I welcome the positive response from the ICT industries to these challenges and their active involvement in the future of the DLM-Forum, for example in the event proposed in Barcelona in May 2002, to coincide with the EU Spanish Presidency.

The information contained in the following pages is one of a series of six ICT Industry White Papers produced by leading industry suppliers, covering the critical areas that need to be addressed to achieve more effective electronic document, records and content management. I am sure that the reader will find this information both relevant and valuable, both as a professional and as a European citizen.

Erkki Liikanen
Member of the Commission for Enterprise and Information Society

Preface Sponsor

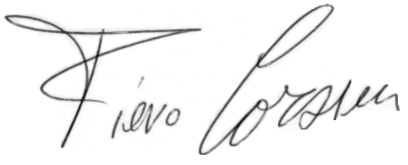
The Internet transformation has just started. Today, less than 5% of the world's population has access to the Internet. In the future, the Internet will be a part of the fabric of mankind - something we will hardly notice. And we will be able to access the Net through a multitude of devices - our cars, watches, and even household appliances.

Due to these changes, government has just entered a transformation process, which will have a larger impact than any other changes that ever happened since the industrial revolution. The way democracy will be practiced and the way public services are delivered will change through widespread access to information and instantaneous interaction between citizens and government.

Applications that seem futuristic today such as e-voting will be commonplace tomorrow. The greatest challenge with this transformation is not lack of vision but the way governments are organised. In the future, governments will find ways of organisation that will be much more user-centric, reshaped through the eyes of the citizens - and available 24 hours a day, 7 days a week.

Accessing the information of today and leveraging yesterday's information for knowledge and insight is at the very core of this transformation. It will require an information robust, secure, scalable and flexible infrastructure based on open standards. User access to information and protection of information are key issues to address to make the transformation happen.

Governments increase the focus on making information available and on involving citizens so they become a more integrated part of governmental processes, and the IT industry develops infrastructures that can help managing information and navigating across it. Collaboration is key to success. Therefore I welcome the initiatives of the DLM-Forum and wish you good reading of this white paper about user access and information protection.



Piero Corsini
Vice President
EMEA Public Sector
IBM

Table of Content

1. Introduction

Information volumes are growing exponentially these years, and IT software and hardware technologies to access the information are developing rapidly. Today, the PC is still the dominating device for user access. In the future, a whole range of mobile devices like mobile phones, pagers, and PDAs will become very commonly used information access devices. These development trends impose challenges related to the individual's ability to extract relevant information, protection of personal information and protection against cyber-attacks.

Through a series of initiatives and best practices within the area, this paper will describe the opportunities and issues related to user access and information protection. The following key topics for user access and information access will be addressed:

Open access to public information: The needs for a) higher transparency in government and b) cost reduction are primary drivers towards open access. Issues regarding litigation, privacy protection and networks attacks need to be addressed in order to provide secure access to citizens.

- Access methods: with increasing volumes of increasingly complex information, the ability to locate and identify relevant information is becoming key - with the portal as a paradigm for the rich function needed for information access.
- Standards: Which standards are relevant to user access and information access?
- Planning for any significant IT application requires knowledge about standards – in particular with open application that will interact with many other systems.
- Protection of public information is not only about how to avoid hacker attacks. Governments need validated audit trails of their information interchange with the citizens, and there is a need for building proof of authenticity into the information infrastructure.
- Trends: User access and information protection is a journey so where do we go from now?

The paper will describe the main drivers for architectural change. The tasks to do in the future are significant but the potential for information access is enormous. I wish you good reading.

John Mancini
AIIM International

2. The challenge of open access

2.1 Information access in Europe

The eEurope action programme was based on the observations that the 4 major reasons for Europe to lag behind US and Canada were

- Slow and expensive Internet access
- Insufficient digitally literate population
- Lack of entrepreneurial culture
- Public Sector not fully exploiting new applications and services

A number of action programmes have been put in place all recognising that in Europe the Public Sector has the potential of acting as a locomotive for the eEurope transition, especially in the area of eProcurement, where public procurement would trigger eBusiness, and in the eContent area where substantial content from the public sector could be used to stimulate new business opportunities and in particular to increase quality of life and skills among the European population.

This again will help social inclusion and provide a basis for a better understanding of our common European cultural heritage.

Seen in a records management perspective, these action programmes can stimulate the citizens' access to public files and allow online transactions directly between citizens and government as well as between business and Government. The pressure on the Public Sector as a result of adverse demographic development simply makes it necessary to look for new and more effective ways to handle Public services.

The recent 'Web based survey on Electronic Public Services', October 2001 (Cap Gemini, Ernest & Young for DG Information Society, October 2001) analyses the huge diversities in the Public Sectors maturity in using ICT and concludes, that "The on-line development of Public Services is enhanced by coordinated service provision and also, that complex administrative procedures require important back-office reorganisations "

2.2 A common framework for information interchange

A coordinated service provision calls for a common framework and common understanding of the basic standards needed for interchanging information, and for describing and delivering services to the public using Internet technology.
The following model seems useful for describing the interaction between the stakeholders in the e-society:

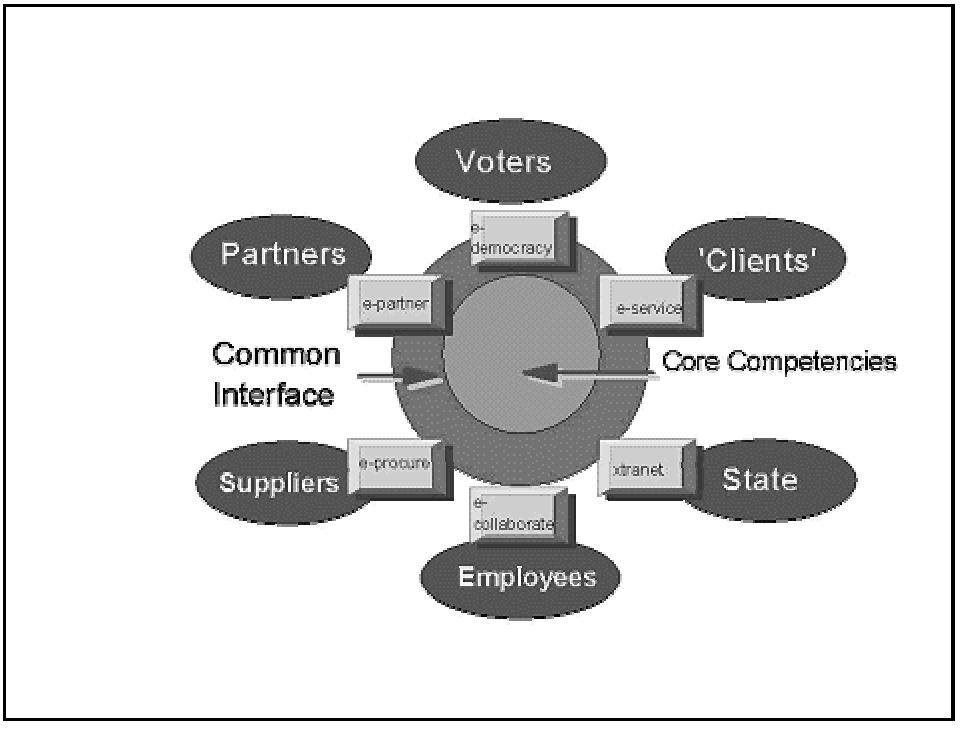


Figure 1: System design

The figure above illustrates that while the application areas differ, depending on which stakeholder we are addressing, the basic idea is to create a system design from the outside in - based on the customer‘s needs for information access and interchange - not on particular departmental views and existing information structures.

In order to accomplish this a common interface model is needed –covering information access, information protection, and information presentation.

2.3 Drivers for open access

The rapid growth of e-commerce has been good for both IT oriented corporations providing solutions and those companies which have been able to shift quickly from the traditional brick and mortar images such as buildings and manufacturing lines and transforming themselves into virtual companies.

Consumer purchasing habits have shifted more easily, naturally and rapidly in some industries than others (books and music vs. groceries and home mortgages). Another paradigm shift is to one of a market economy, were goods and services are exchanged to one of Intellectual Property where concepts, ideas, and images will be exchanged via these networks.

These drivers are pushing technology development in general, and the governmental sector is benefiting from them.

2.4 Exposure to Litigation

In some societies today (e.g. the United States) people settle many differences in court. In addition, government regulations govern many aspects of business in any country. Therefore, the importance of good records management and information retention policies has become a critical element in business today both with the government agencies and corporate entities.

e-Commerce results in e-Contracts and with them the following questions:

- What is the Viability of the transaction (explicit or implicit)?
- Where, how, and when was the transaction authenticated?
- What is the evidence that the transaction occurred?
- What is regarded as the record for the transaction (agreement of both parties, to include term and conditions, and is there evidence of the obligation)?

2.5 Privacy-protection versus behaviour tracking

As Open Access typically will take place using an electronic portal as the control point, most portal solutions on the market put a lot of emphasis in being able to track customer/consumer behaviour as a part of their CRM strategy; while this in some geographies may be legal, in other countries such as the European Union Countries, the protection of consumers and individuals privacy in most cases make it illegal to use these tools unless the consumer specifically agrees to this 'service'.

The more modern type of Portal solutions in response have created solutions for 'anonymous' users to log in as guests and in this way be able to track behaviour for the benefit of designing the solution but without relating the information to an individual.
This solution will not work for citizens accessing public databases containing personal and private information about themselves and all other citizens. These databases have to be kept securely behind double firewalls and are typically subject to special audit regarding the procedures for back up/recovery and of course access rights.

Using Digital Certificates will open up the opportunity for individuals to securely access and even change their own data and then by storing it in an encrypted form, only make it available to the case officer and themselves. Theoretically a number of public databases could even 'change ownership': letting the citizens own their own data and be responsible for their validity.

Digital certificates make it possible to track and log the changes and the behaviour of the owners on the web site in question. This log can be stored as securely as the personal data and only be opened by proper decision by a court. In this way the digital certificate can help protect against 'big brother' spying, eavesdropping or against revealing personal information by mistake.

Another aspect of behaviour tracking, which will emerge when portal services are expanding is the technical infrastructure needed to manage secure access for an individual to different applications residing in different domains. The possibility for providing a single point-of-control where the access rights can be maintained on a cross-domain level is of growing importance. Establishing a single point-of-control where the access policies and rights are maintained would be of great benefit for the citizens that would otherwise need to log on to a new system every time they required access to another domain.

And respectively, for the administrators of the different systems, that will have one access point to check and update access rights without having to program this into every new application. This solution is called Policy Director and has been widely accepted where multiple domains are accessed through same portal server.

2.6 Digital rights protection

Especially for the Media industry protection of digital rights play an important role as their income over time will be eroded if the assets are freely copied using Napster-like technologies. Also the Public Sector has similar needs or because of the value of the information like maps, music, patent information or legal information of various kinds. On top of this, use of restricted and classified information may need protection in a similar way.

The IT industry and the Media companies have carried out a great deal of research in different technologies over time – watermarking, encryption, records management – and this has resulted in a series of software solutions to manage the digital rights regardless of the format of the digital assets. Within EU, under the auspices of the INFO2000 programme, the MMRCS (Multimedia Rights Clearance Systems) project has examined technical, legal, and business issues surrounding the development of digital rights clearance systems. Such clearance systems are seen as a possible infrastructure for exchanging intellectual property, ensuring safe and reliable payment, controlling production of copies etc.

The following functionalities would be required in a comprehensive media management/rights management system:

The challenge is to create an e-commerce software solution for digital distribution of media. A set of software products to build a technical platform with tools and security features that enable copyright owners and market intermediaries to conduct commerce and deliver electronic media content over various types of digital transmission systems.

The solution should include a set of content mastering tools, a content host and distribution subsystem, a clearinghouse for security, rights management, reporting and payment interfacing, an electronic store subsystem and a ideally also some tools to assist developing PC client software for security, rights, library and device management

The requirements for a Content Mastering Program is that it will assist content owners to specify business rules of use and to selectivity perform pre-processing, encoding, watermarking and encryption of their creative works. These digital assets should also be packaged into secure containers for electronic distribution to content hosting sites.

The requirements for a Clearinghouse solution are that in contains functions for managing security, rights management, reporting, and payment interfacing. A Clearinghouse should be the central control point for managing and authorising transactions. It should verify licensing requests; issue licenses that enable consumers to access content and it should provide information to facilitate royalty payments. Optionally, a Clearinghouse should offer credit card transaction processing.

3. Accessing Public Information

3.1 The portal as a paradigm

During the recent two years portals have become the paradigm for access – not only to information but also to applications and knowledge.

To some, portals represent communities; to others, they are trading hubs or e-marketplaces; and to many, they are integrated desktop environments. Putting aside the hype, a common theme of substance underlies portals – a theme of greater levels of integration.

From a unifying technology perspective, a portal is a single integrated point of comprehensive, ubiquitous, and useful access to information (data), applications, and people. This definition encompasses all the different views of the purpose and functionality of portals. But more importantly, strong pursuit of satisfying this portal definition will help evolve the next generation of integrated services and business processes.

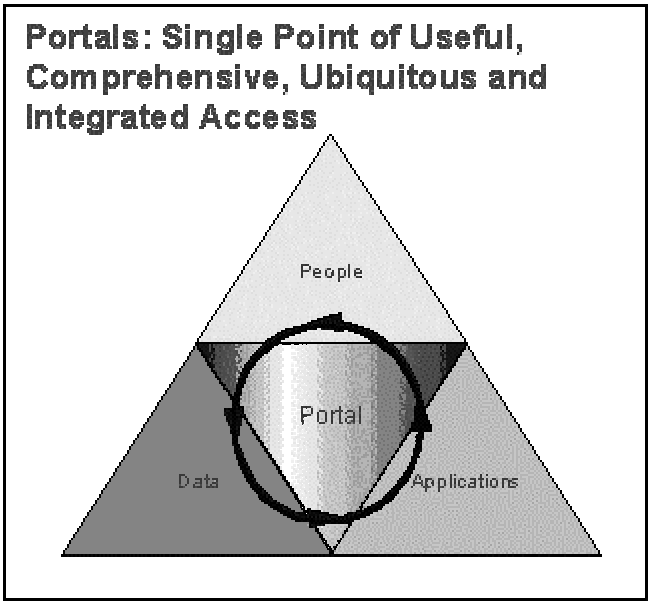


Figure 2: Portals

3.2 Usage scenarios - corporate, personal, marketplace portals ...

There are many different types of portals, but at a high level all portals can be viewed as an instance or a combination of one of the following types.

- Corporate portals or corporate desktops provide employees with access to information and applications. One can envision that such portals will evolve to become the next

generation of integrated and intelligent desktops that can be accessed from anywhere and from any device. Many information-centric applications such as knowledge management, business intelligence, and customer-relationship management software are being integrated and deployed via similar portals.

- Inter-organisational portals are owned and managed by a single organisation or enterprise but support business processes such as supply-chain solutions and procurement among customers, partners, and vendors across different organisations. Supply-chain portals aggregate parts, inventory, and pricing information from a large number of suppliers, support interaction, collaboration, and dynamic partnerships and help manage/coordinate the end-to-end flow of a business process.
- e-marketplaces are portals or trading hubs that connect buyers and sellers in virtual marketplaces. Such portals are cross-enterprise and can cater to specific industries such as the steel industry or semiconductor industry. In addition to an infrastructure that is common to most portals, e-marketplaces may need specific service -- for example, bidding and auction services.
- Personal portals, provided by the likes of Yahoo, Excite, and Netscape, provide users with a customised, first point of access to the Web. The business intent is to grow the portal's user base by providing useful information and functions that in turn drive advertisement revenues. Community portals are a variant that cater to specific communities of interest by providing tailored content.
- ASP portals provide integrated access to data and applications and are supplemented by a set of additional services to support the business models of Application Service Providers, or ASPs. The new business model of ASPs expects renting of hosted applications to be a viable and economic alternative for users.
- Portals are being considered for the next generation of integrated application development platforms. Instead of disconnected pieces of applications that must be sequenced manually, application vendors are exploiting the power of the integrated, process-centric view of the design and development process. Common services for almost all design and development efforts should include:
    - Access to information– structured (aka data) or unstructured (aka content)
    - Collaboration between parties both within and outside the enterprise
    - Information flow between applications, application to people, and people to people.

Many envision significant opportunities for growth in Web-based re-deployment of existing applications that tightly integrates the above components under a portal environment.

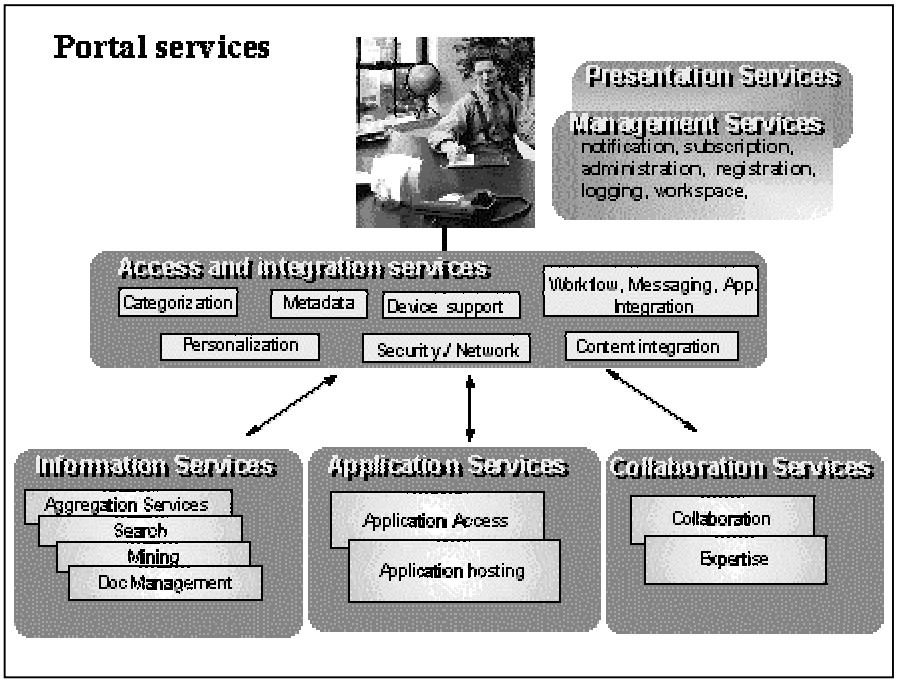


Figure 3: Diagram of portal services being covered in the following sections

3.3 Information aggregation

A portal must provide aggregation services giving users a single point of access to multiple heterogeneous data sources, both structured and unstructured – including relational database, multidimensional databases, document management systems, ERP databases such as SAP, e-mail systems, web content, Web servers, news feeds, and various file systems and servers – in a transparent manner.

The system must support all types and formats of content – audio, video, image, text, postscript, and so on. The aggregation services connect all these data sources in a federated manner across multiple locations and companies. From a user's perspective, access must be transparent and seamless. The user should not need to be aware of the exact location or nature of the repositories, and all information types should be accessible through a common federated search and retrieve mechanism.

This notion is by some being described as a “Virtual Repository” – information may reside in many different places but to the user it looks like one large single repository.

3.4 Web content

Portals need to be supported by a web content management platform consisting of well-defined processes and tools that support the collaborative activities of content creators and site administrators. These activities include web based life-cycle management, authoring and designing Web based assets, controlling content delivery to the user and increasing value through appropriate content re-use and personalisation. The current 1st paragraph now becomes the 2nd paragraph and the lead-in sentence to the bullet list should read: To maintain, manage, and provide access to various document types, portals also require an integrated set of document management services including:

- Document check in and out
- Document version control
- Document security for specific users or groups of users
- Extensive audit trails to identify who did what to which document and when it was done
- Lifecycle management of documents, including retention management and permanent archiving/destruction

3.5 Search and mining

Search technology helps users filter a growing morass of information to find information that is useful and relevant. Many different search techniques exist, but elaborate discussion on such techniques is beyond the scope of this paper. However, one important search technique is parametric search, which associates name/attribute pairs to search keywords. This technique is especially relevant to electronic commerce where, for example, a customer needs to find dress pants that are black, size 34, and cost under $40.

All search techniques rely on meta-data to satisfy a search request. Many search infrastructures rely on Web crawlers to create the search meta-data. A Web crawler retrieves Web pages for use by a search engine by recursively scanning and extracting URLs from a starting document or Web page. Every page that is scanned is indexed, sometimes ranked, summarised, and analysed for its contents – the information stored as meta-data. Due to the huge and growing number of pages on the Web, and the constantly changing nature of some pages, Web crawling can be a serious challenge and resource drain. Meta-data can also be created using other techniques or directly imported from data warehouses

3.5.1 Sample scenario: Searching for citizen's information

The Dept of Social Security has an extensive collection of photographs, claims, policies, assessor's notes, reports from experts, and other process- related documents. The Department keeps all memos that are sent to claimants, along with medical and appraisal electronic forms in an e-mail system. The Department archives all assessments, declarations, notices, and statements in a report repository for long-term storage and quick access. The Department stores all claim forms, photographs, and letters received from

claimants in a content management system running on a legacy operating system. The Department keeps reports from experts in a relational database Data Warehouse. The Department also stores media assets such as high-resolution graphics and training videos in an other content management system for the advertising, public relations, and education departments to share. In addition, the Department keeps information, such as company procedures, on its department intranet.

Claims, claimant calls, and general claimant servicing cannot be handled with the content from one system because employees need to access all customer information. To provide claimant service, employees require simultaneous access to a variety of content systems, some computerised, some manually maintained. The Department needs a solution that connects their computer systems, manual filing processes and their department intranet for searching and retrieving information. They also want to expand their use of workflow processing.

Many different employees need to access documents, from clerks to claim assessors to field agents. The Department must restrict access to certain items, while providing unlimited access to others. The Department also wants an easy-to-use interface to reduce the need for training of staff and to enable citizens to access their own data as they move to self-service electronic delivery systems.

3.5.2 Information mining – structuring the unstructured

Information mining provides linguistic services to find hidden information in text documents on content servers. During processing of the text documents, metadata (information about data) is created that can be summarised, categorized, and searched.

The need for tools to deal with online documents is already large (we need only mention the internet) and is growing larger. Consultants have predicted that unstructured data (such as text) will become the predominant data type stored online. This implies a huge opportunity: to make more effective use of repositories of correspondence and other unstructured data, by using computer analysis.

But the problem with text is that it is not designed to be used by computers. Unlike the tabular information typically stored in databases today, documents have only limited internal structure, if any. Furthermore, the important information they contain is not explicit but is implicit: buried in the text. Hence the “mining” metaphor – the computer rediscovers information that was encoded in the text by its author.

Many of the tools used in text mining can be seen as information extractors that enrich documents with information about their contents. This information can be used as metadata about the documents. Metadata is structured data that can be stored in a database and could be used in turn for Data Mining. This process of annotating documents with metadata is shown schematically in figure 4.

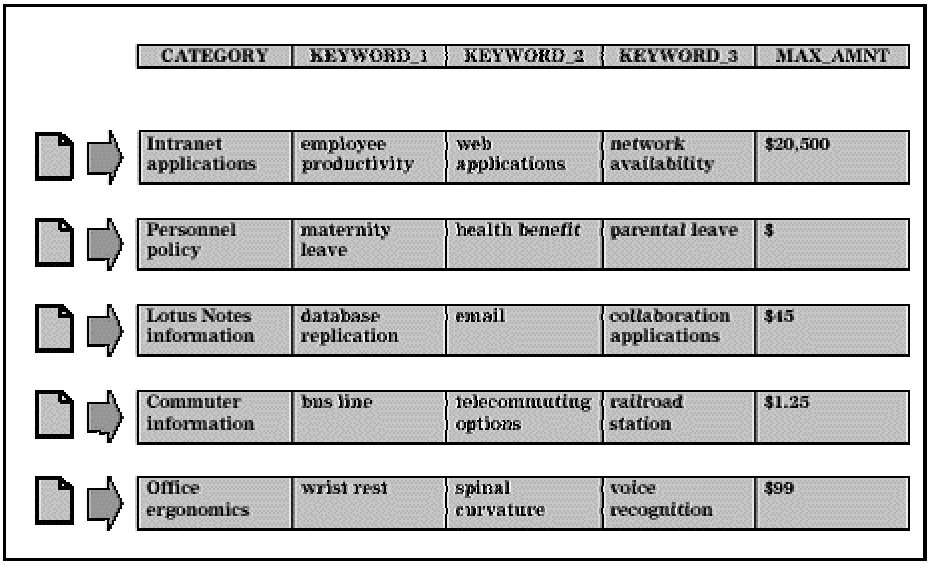


Figure 4: Annotating documents with metadata.

Schematic showing how tools can be used to populate database rows with information about documents – in this case the categories in a cataloguing scheme they belong to, the three most significant technical terms in each document, and the largest money amount they refer to.

Often the first step is to extract key features from texts, to act as "handles" in further processing. Examples of features are the language of the text, or company names or dates mentioned. After feature extraction, the next step may be to assign the documents to subjects from a cataloguing scheme, then to index them for ad-hoc searching. The tools that support these and other functions will be described below.

3.6 Text analysis functions

3.6.1 Language Identification

A language identification tool can automatically discover the language(s) in which a document is written. It uses clues in the document's contents to identify the languages, and if the document is written in two languages it determines the approximate proportion of each one. The determination is based on a set of training documents in the languages.

The accuracy in the tool suite is usually close to 100% even for short text. The tool can often be extended to cover additional languages or it can be trained for a completely new set of languages. Its accuracy in this case can be easily higher than 90%, even with limited training data.
Applications of a language identification tool include: automating the process of organising collections of indexable data by language; restricting search results by language; or routing documents to language translators.

3.6.2 Feature Extraction

Feature extraction recognises significant vocabulary items in text. The process can be fully automatic – the vocabulary is not predefined. Nevertheless the names and other multiword terms that are found are of high quality and in fact correspond closely to the characteristic vocabulary used in the domain of the documents being analysed. In fact what is found is to a large degree the vocabulary in which concepts occurring in the collection are expressed. This makes Feature Extraction a powerful Text Mining technique. Among the features automatically recognised are

- Names, of people, organisations and places
- Multiword terms
- Abbreviations
- Other vocabulary, such as dates and currency amounts

3.6.3 Clustering

Clustering is a fully automatic process, which divides a collection of documents into groups. The documents in each group are similar to each other in some way. When the content of documents is used as the basis of the clustering, the different groups correspond to different topics or themes that are discussed in the collection. Thus, clustering is a way to find out what the collection contains. To help to identify the topic of a group, the clustering tool identifies a list of terms or words, which are common in the documents in the group.

Clustering can also be done with respect to combinations of the properties of documents, such as their length, cost, date, etc. An example of clustering would be to analyse e-mail from citizens to find out if there are some common themes that have been overlooked. The effect of clustering is to segment a document collection into subsets (the clusters) within which the documents are similar in that they tend to have some common features.

Clustering can be used to

- Provide an overview of the contents of a large document collection
- Identify hidden similarities
- Ease the process of browsing to find similar or related information.

3.6.4 Categorization

Categorization tools assign documents to pre-existing categories, sometimes called “topics” or “themes”. The categories are chosen to match the intended use of the collection. By assigning documents to categories, the tool can help to organise them. While categorization cannot take the place of the kind of cataloguing that a librarian can do, it provides a much less expensive alternative.

Applications of categorization include:

Organising intranet documents:
For example, documents on an intranet might be divided into categories like “Personnel policy”, or “Commuter information”
It has been estimated that it costs at least $25 to have a librarian catalogue an item. It is clearly impractical to catalogue the million or so documents on a large intranet in this way. By using automatic categorization, documents can be assigned to an organisation scheme, which makes it easier to find them by browsing, or by restricting the scope of a text search.

Assigning documents to folders:
Categorization can also help to file documents in a smarter way. For example, it could help a person who has to assign e-mail to one of a set of folders, by suggesting which folders should be considered. While the actual categorization of documents is fast and inexpensive using automatic tools, the definition of the categorization scheme requires some care.

3.7 Five Examples of the use of Text Mining

Making your marketing more effective
You are a car manufacturer. You want to know which audiences you should target with your marketing campaigns, and find out who your main competitors are. From a large collection of car-related documents, including newspaper articles and reviews, use a Feature Extraction tool to extract the makes and models of cars.
Then, using Clustering tools, identify those makes and models most often discussed within a group. The group, or cluster, would have the names of the makes and the models that are typically viewed as competitors. Also, the name of the source can lead you to where your car was discussed. The demographics for the magazine where the information was discovered might lead you to purchase advertising space in that magazine.

Show me more like this
When searching for information in huge document collections such as the World Wide Web, it is often difficult to construct queries that reliably find documents that match your information needs. Text mining technologies can help to improve this in various ways. The Text Search Engine may have a function that allows you to “show me more like this”. This is also known as a narrow query. It works as follows: suppose you issue a vague query and are browsing the documents in the results list. You might find a document that discusses exactly what you are interested in. You then need a function that can return similar documents. You rate this document as being highly relevant to your area of interest. The “narrow query” function turns this highly rated document into a query, which is then submitted to a search engine.
The tools in a text mining application can be used to do this efficiently. The first step is to use the Feature Extraction tool to build a vocabulary dictionary for the collection being searched. Then, each document is pre-analysed and its most significant vocabulary items

are stored in a database. These items are then used as the query sent to the Text Search Engine when the "narrow query" function is started.

Searching with categories

In text search, one of the main problems is that many of the documents near the top of a results list are not relevant to the query. For example, a person looking for vegetarian recipes might search for "eggplant" and find not only recipes for cooking eggplants, but also documents about growing eggplants. One way to avoid this is to have the user select some categories or information that they want to restrict their query to.

By specifying that the category of interest is "cooking", they can see a more focussed results list. First the Text Search Engine is used to find documents that contain the term "eggplant".

Then the Topic Categorization tool sorts these found documents into predefined categories, among which is the cooking category. The documents in this category are presented to the user as a search result.

Helping the search engine to read your mind

After a search using the Text Search Engine, an application could allow you to mark the relevant documents in the result list. The marked documents could be passed to the Feature Extraction tool to isolate the terminology used in them. The extraction could be used to reformulate the query by adding important terms from relevant documents, and excluding important terms from irrelevant documents.

The result of that query would then be more focused on documents related to your information interest. That is, the precision of the retrieval would be improved.

Surviving a flood of e-mail

You sell goods from a Web site and display your e-mail address on the home page. Because it is relatively easy for customers to send e-mail, the volume is much larger than from conventional mail. Currently, you employ someone to read each item of mail and forward it manually to the correct recipient. The process is slow and expensive. The e-mail also contains valuable customer feedback about your Web site, as well as comments and questions about the products and services offered there.

So, you would like to use this information to improve your business. You can use a Topic Categorization tool as the basis of an application to do the routing. Collect a sample set of messages for each recipient and use these as training data for the Topic Categorization tool. Then, for each new message the Topic Categorization tool suggests a list of recipients, with confidence scores. Only if none of the confidence scores are above some threshold value, meaning that the Topic Categorization tool could not find enough evidence in the message to make a decision, does the message have to be read by a person. With a system of this kind, the messages can be routed automatically to the correct destination. You process them quickly and at low cost.

But now consider how much more information is in the e-mail messages. Even those containing complaints have valuable information about your Web site and the company's products and services.

To make the most of this information, you could mine the archive of e-mail messages. Using the Clustering tools, you can group messages together by similarity, and store the

document-cluster relations and the keywords that characterise each cluster. With these in place, and using the Text Search Engine's search facilities, the archive becomes a tool for finding and analysing common customer concerns.

3.8 Consolidated and Syndicated Content

To dynamically assemble personalised documents, portals require content integration tools to manage discrete content objects, which can be combined on the fly to create different target formats. The target format may be PDF, HTML, CD-ROM, or print media. Language translation of discrete content objects can be handled automatically during the assembly of the target document. By providing the appropriate granularity level, a content management system can easily generate the necessary target documents without having to maintain redundant information.

Portals must also provision syndicated content from a wide range of leading content providers. Such content includes news headlines and feeds, weather, stock quotes, financial and vertical industry news, as well as sports and entertainment news.

3.9 Metadata

Metadata or information about data is very important when it comes to handling unstructured data such as text.

The Public Sector has a very special need for standardising the descriptions of the huge amount of textual information that today resides in a multitude of data bases and archives in various, decentralised and geographically dispersed institutions.

One of the most obvious obstacles to using the public sector's eContent is just this stovepipe-like organisation of work along with processes and data.
In some countries standardisation work has begun to create metadata based on XML so that a common database of 'pointers' can help to solve some part of the problems of inter-organisational barriers.

A Meta-database is a sort of 'Catalogue' of the related information and can be regarded as a portal solution seen from the various content owners and users' point-of-view. The mere creation of the Meta-database, however, does not solve the problem of access, and the vision is therefore, at the same time as creating the Catalogue, the related access rules and associated services (and, where possible) the transaction types or services can be described using for instance the UDDI-protocol (Universal Description, Discovery and Integration) together with WSDL – Web Service Definition Language.

In this way the 'Catalogue' becomes an enriched catalogue that can then grow into a 'market place' for (in this case) government related services around the data and not just a guide to find the right information. Further, the mechanism for users to access these services will be enhanced by using SOAP – Simple Objects Access Protocol.

The vision behind the very large consortium that are backing these standards is to create an interoperable, global catalogue for services and data that can be accessed by anybody that is authorised from anywhere in the World.

3.10 The role of XML

The promise of using XML as the common data structure is that it will be technologically neutral allowing different types of servers, displays, and communication tools to benefit from the ever-increasing richness of information.

For one of the most aggressive and progressive organisations pursuing this, see http://xml.apache.org, which has the goals to create commercial-quality XML standards and XML-solutions developed using Open Source methods and to provide feedback to the standards bodies, IETF and W3C.

Many standards are of key importance to user access - including HTTP, SSL, XML, LDAP, TCP/IP, Java, EJB, SQL, ODBC, JDBC, and CWMI. To support integration of a diverse set of data, applications and processes, eXtensible Mark-up Language (XML) is a key enabler.

From a content management point of view, XML provides a flexible source data format for Web delivery and content integration, permitting multiple presentation formats and/or media types (Web, CD, paper) to be supported from a single source repository. XML allows smaller fragments of text, graphics, etc. to be managed as reusable information objects that can be dynamically assembled into final form documents or Web pages.

Flexible XML-based query language (XQL) and database schema standards such as XML-schema will enable a common query infrastructure for search. Also, XML based meta-data standards allow flexible interchange of document meta-data across systems. Such interchange supports federated searches as well as integration of document meta-data. CWMI, for example, enables inter-operability with other systems and extensibility of stored information.

XML enables interchange of information across complex, application-specific business-to-business e-commerce applications (for example, airline technical data and semiconductor data sheets). XML enables application integration as well as integration of data into applications and processes.

At the client level, XML affords rich presentation formats that can be individually tailored and controlled by the information consumer.

For pervasive devices, XML-based mark-up languages will help transcode content from one form to another. VxML, for example, will enable integrated speech support in portals.

3.11 The enterprise content management challenge

In the blink of an eye, in parallel with e-business, e-government has evolved from futuristic vision to revolutionary frenzy to a practical reality of real business for every organisation, every day. One lesson learned from the first is that practical success demands a solid IT infrastructure, a long-lived foundation capable of supporting not only today's mode of operation but new models of interaction and new applications yet unknown, throughout the organisation and across organisations.

The fuel of e-government is information, both "structured" – usually called data – and "unstructured" – referred to as content. The content that drives e-government is more than simply static web pages. It also includes:

- Dynamic web content, data in relational databases, personalised for each web user
- Documents that support back office processes and inform citizens, organisations, partners, and employees – from reports to e-mails
- Rich media – digital audio and video – rapidly transforming not only the entertainment industry but training, education, marketing, and customer relationship management in every industry including government.

Today, the web can make all of this content immediately available, while hiding differences in underlying content formats. You click on a link, and the content displays (or "plays") in a browser, whether it's an HTML page, a word processing document, a scanned image, a mainframe report, or a video clip.

Thus, in order to effectively share and distribute information across the enterprise and beyond – to employees, citizens, organisations partners – e-government today demands a unified framework for managing, web-enabling, and personalising delivery of all these disparate forms of content. This framework is called Enterprise Content Management (ECM).

It is a mistake to equate "content management" for e-government with simply web content management, which typically ignores both documents and rich digital media. Today's web content management system technologies do a good job (some better than others) at managing version control and publishing to web sites during the active life of the information.

However, most of the popular WCM solutions do not address the records retention (inactive life of information) aspects of web content.

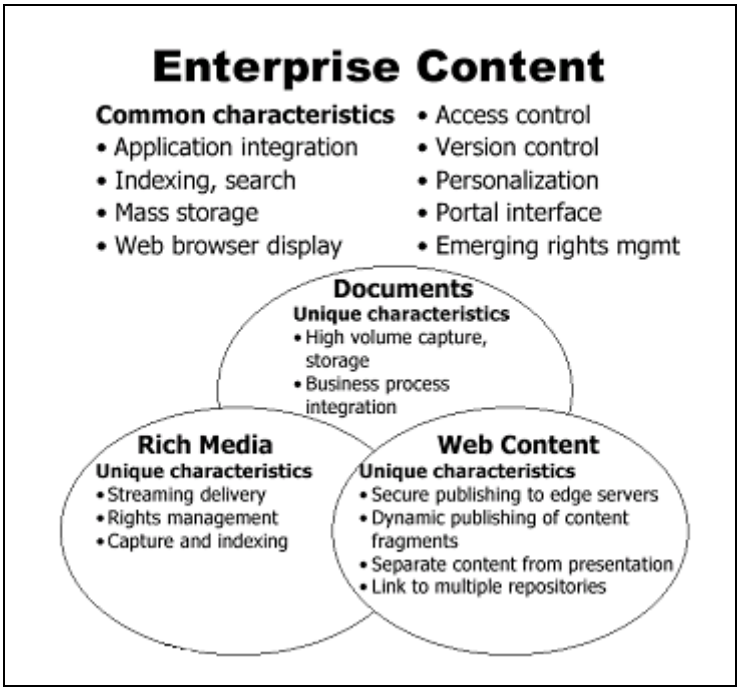


Figure 5: Enterprise Content Management characteristics

Enterprise Content Management embraces three historically separate technologies: web content management, document management, and digital media asset management. While outwardly dissimilar, all of these forms of enterprise content share similar needs for mass storage, search and access, personalisation, integration with legacy applications, access and version control, and rapid delivery over the internet.

This commonality suggests that rather than separate point solutions for each type of enterprise content, enterprise content management can – and should – be implemented on top of a set of unified ECM components, a new extension of the e-business infrastructure. An infrastructure approach to ECM allows familiar economies of scale: Applications across the organisation can leverage common platforms and peripherals, lowering the total cost of ownership – not only hardware and software, but system administration, training, and custom development as well. New e-government initiatives can cost-effectively take advantage of the full range of enterprise content.

Legacy content can be easily published to web sites, and via portals enabled for secure, personalised delivery to citizens, partners, organisations and employees across and beyond the organisation. ECM infrastructure also allows emerging technologies like digital rights management and XML web services to be implemented consistently and cost-effectively across all forms of enterprise content.

What then makes a set of ECM components an infrastructure?

- Information integration
  The most important factor is the ability of any business application to access all forms of enterprise content – from business documents to rich media to dynamic web content – stored in any repository, through a common software interface. In fact, customers want to be able to do federated search – a single query retrieving content from a heterogeneous set of repositories distributed on the network – and they want to use federated indexing data for personalising that content in portals and for e-commerce.
- Repository scalability and robustness
  Content repositories must be able to scale cost-effectively from small departmental solutions to enterprise-wide applications used by thousands of employees, to customer-facing e-business sites getting millions of web hits per day. Scalability also means ability to manage huge content volumes, particularly important for documents and rich media, where repositories may house hundreds of gigabytes of content yet must maintain quick response times. And it means the ability to cache and distribute content to the "edge of the network" to optimise delivery speed, yet maintain tight central management control.
- Openness
  ECM infrastructure must not just be based on open standards, but must expose its functionality to any application – including those of competing vendors – supporting a set of published application program interfaces (APIs). ECM infrastructure should support leading infrastructure components from more than one vendor, including server platforms, database management systems, content repositories, and packaged applications.

3.12 Presentation support

The presentation of data to the end-user has to be customised based on either a pre-determined set of criteria or the user's perceived needs based on historical data collected from previous browsing sessions. It is becoming increasingly common to deploy the end-user interface via a Portal.

A Portal is composed of a set of portlets (or mini windows) usually chosen explicitly by the user from the total set available or a subset based on the user's profile.

A portal must be supported by a set of presentation services that manages the look and feel of the portal interface and enhances the user experience. Customisation of a portal occurs at various levels.

Users may wish to tailor the look and feel of the user interface including the taxonomy of applications that are integrated into the portal desktop. Users may desire a different look and feel from a specific portal based on the task at hand. For example, the user interface for recreational browsing at home may be different than the one used at work for business purposes.

Portals offer a desktop look and feel where the browser is partitioned into frames and real estate carved out and allocated to applications. Application and content sources are wrapped into components (commonly referred to as portlets) that can be positioned on the screen in a plug and play manner.

A portal desktop will need to support task management, which is not intrinsically supported by the current generation of Web browsers. Such task management may be extended to pervasive devices as well.

Portals must support access to data, applications, and people via a variety of devices -- including phone (wired and wireless), pagers, fax, handheld devices, laptops, desktops, workstations, and servers. Transcoding services are required to filter or convert content to match the form factor and capabilities of the target device.

Alternatively, appropriate type of content may be selected to best suit a particular device type. Services for disconnected use help users transmit and receive information reliably when they establish a network connection.

3.13 Application Services

Application services for portals support two major functions -- application access services and application hosting services. Application access services are supported by Web application servers, which provide seamless access to both Web-enabled and legacy applications.

Access to legacy applications is supported by a rich set of connectors running on the Web application server. Application hosting services simplify management and deployment of applications that are hosted externally by an ASP.

3.14 Collaboration

Portals may need to support real-time collaboration so that it is easy for users to find colleagues, partners, customers, and suppliers online, and communicate with them regardless of where they're working. These services allow users to chat, share live documents and applications, and create an instant-shared workspace where team members can centrally communicate.

Mail and shared calendar services are also instances of collaboration tools that are increasingly finding their place as integrated portal services. A unified messaging system for fax, mail, and telephone that converts text to speech is another growing portal requirement.

Another useful collaboration service for some portal applications is expertise location, which maintains profiles that can be queried directly by users to locate experts by skill, experience, project, education, job type, and many other attributes.

The profiles get loaded through a variety of measures. Skill data is acquired through a metrics tool in the discovery engine, which mines topics out of documents and cross-references them with user activity to infer interest and skill.

3.15 Personalisation, strategies and tools

Personalisation encompasses the ability to provide the user with the right content both from the user's and Web site owner's perspective. It is based on gathering information about the user or communities of users and delivering the right content at the right time based on current context.

A personalisation algorithm determines whether content is presented to the user, and if so, in what order of priority. There are several different types of personalisation techniques:

- Profile based, where user information is profiled in a directory and used for making recommendations. The information may be provided by the user or set on the user's behalf.
- Collaborative filtering is a personalisation scheme to effectively identify individuals who share similar tastes. Users are asked to rank alternatives; a rank order of matches with other users determines proper recommendations.
- Real-time recommendation learns and predicts a user's preferences by observing usage patterns and from rankings obtained by methods similar to collaborative filtering.
- Content-based filtering is a scheme for automated personalisation where suitability of a document is based on analysis of keywords and can be used when the selection criteria is not subjective.

4. Protecting Public Information

4.1 Security issues as market drivers

Information Privacy and Security are major issues in today's information based society and economy. To illustrate the significance and importance of this fact the following facts are provided: In 1996, 2,573 reported security break-ins on web-based systems took place –compared to 15,157 just in the first 3 quarters of 2001.

As the percentage of growth increases, it becomes imperative that these corporate assets/records be protected. Compromised bank transactions and social security numbers is an inhibitor to any large-scale deployment of open access.

4.2 Management vs. Retention

In today's organisational environment we have an all-to-real dichotomy of management versus retention.

On the one hand we have legal mandates, regulatory requirements, and good business practices which dictate that e.g. e-mail must be retained in accordance with established retention requirements.

On the other hand, we have organisational policies requiring the deletion of large in-boxes and files after a specified length of time or upon reaching space limits set within the system.

When employees encounter this dichotomy they will commonly react in one of the following three ways:

- Either save the attachments and message to their local hard drive, thereby losing control over retention requirements.
- Or print and file paper copies, increasing hardcopy storage requirements, which is inefficient and results in loss of control over retention requirements.
- Or simply deleting critical business information - creating a significant retention problem.

4.2.1 Policy management

Electronic Record Management is the key to information preservation and protection. Fundamental to records management is security, personal (privacy) data protection. These prevent cyber crime and unauthorised access to and publishing of information.

The leading security technology in this arena is Cryptography using symmetric or asymmetric keys. In Europe the main public emphasis is on the asymmetric public key encryption (PKI – Public Key Infrastructure.).

4.2.2 Records management - governmental drivers

Some of the drivers impacting the need for improved electronic records management today include:

- Antitrust legislation to ensure that prudent competition is being enforced.
- Privacy Protection – how much control should web surfers have over their personal data?
- Cyber crime – obtaining personal, financial information, hacking into websites, controlling the promulgation of pornography and cyber stalking.
- Sales tax for online purchases – whether or not to levy taxes on purchases made over the web, what are the rules, and to whom do they apply.
- Online Piracy – and the need to control copyright material available on the internet

Some international efforts include „Developing guidelines for conduct of e-business worldwide". This e-Commerce code of conduct is intended to be voluntary. In addition, the countries involved are evaluating consumer protection laws of each country, which would apply to e-Commerce transactions.

And finally, they are working on educating consumers regarding what security and integrity things they should look for when shopping online. The goal of these efforts is to build confidence in the global e-marketplace.

Electronic Records management supports the ability to preserve valuable organisational knowledge, and it improves decision-making by retaining information related to past decisions. It controls the destruction of expired information, and it links paper records with their electronic counterparts.

Records Management technology manages information and records throughout their life cycle (creation, active life, inactive life and archival or destruction).

4.3 Emergence of virtual documents and virtual records

With the emergence of virtual documents and virtual records in today's world of e-Commerce a single object (e.g., hypertext/hyperobject) may not be a record. However, when combined with other objects it may constitute a record.

Emerging from this is the notion of a virtual document or virtual record – a "super" record, consisting of parts from various sources. Web links that provide the connection that creates these "super" records must be preserved. This is critical to demonstrating the integrity and reliability of the "super" record. To demonstrate the need for this "super" record capability, consider the following: Only 66% of web addresses lead to live sites. Many critical website references have already been lost.

Today's sophisticated e-business transactions bring with them an additional challenge. That challenge is managing the three different elements, which comprise this hybrid virtual e-record:

- Content – the data/information contained within the object;
- Context – the circumstances of the creation of the object; and
- Structure – the logical and physical attributes of the object (visual and interactive presentation of the transaction).

4.4 Audit trails

e-Record solutions must have the ability to track and record user and component transactions. User transactions should include: Logon, Failed Logon, Logoff, and Life Cycle Management Events.

Component Actions should include: Add, Delete, Update, View Move, Copy and Life Cycle Transaction.

4.5 Transaction integrity through electronic signatures

One way to ensure the integrity of electronic transactions is via the use of digital (see below) signature technology. The "Electronic Signatures in Global and National Commerce Act", establishes a framework for e-commerce throughout the United States and many other countries via the use of digital signatures.

Advances in the area of electronic signatures have grown a great deal in a short a short time. This is not only a question of the length of the keys used for encryption, but also due to the increased sophistication of encryption algorithms as well as the administrative procedures and mechanisms that will ensure an effective and yet transparent security infrastructure.

Following the European Electronic Signature Directive (in this directive the differentiation between digital and electronic signature is described.) in 1999 and the national legislations on use of electronic signatures that have been accepted by almost all EU countries since then, it was expected that the use of Electronic Qualified Signatures, especially for eGovernment applications would have been widely used; as this does not seem to be the case it is mandatory to investigate what seems to be the reason for this and what could possibly be done to stimulate the use of digital signatures.

This almost seems like a chicken and egg-type of problem: As there are no certificates circulating, not many applications are developed to use them, and as not many applications are available, the time effort – and in some cases also the money required to obtain the certificates, seem to be out of line with the potential benefit one can expect to get from using it.

Especially following the acts of terror on September 11, an increasing pressure has been put on parliaments all around the World to allow surveillance of individuals to a much larger degree than traditionally accepted in modern democracies, and it seems as the use of a digital identity also of this reason might be a safeguard to protect the rights of the individual and at the same time provide a means for the courts to track individual behaviour only after proper juridical decision has been made to reveal otherwise privately and protected information.

In other words, by issuing digital certificates and identities to everybody we can keep our current legislation on protection on privacy without allowing for a Big-Brother-Watching-You type of society.

As this is a political argument to accelerate the issuance of digital certificates to everybody, there is a very rational other reason why Governments should have a huge interest in distributing digital certificates: The results of pilot projects indicate that if you replace standard paper forms with electronic, intelligent forms and use digital certificates to ensure authenticity, administrative savings of between 20 and 50% of case handling and customer related services can be obtained in the public administration.

The bottlenecks in distributing Digital Certificates so far also concerns the media that are being used; if we take the Finnish example, the use of a smart card as a certificate 'container' implies cost at least in the magnitude of 25$ pr. Citizen plus the cost of installing a smart card reader on his/her home PC – and the trouble to install a driver. Add to this that a certificate has to be renewed every 2 years.

This and other similar projects have proved that a 'market approach' will not generate a very large number of circulating certificates. There is simply no financial justification for the citizens to acquire a smart card under these circumstances.

At the same time a substantial number of Europeans are quite used to Home Banking. Home Banking is normally based on soft certificates, where the customer receives a pin code, logs on to the bank and receives a java applet that will generate a key pair on his/her PC and send the public key back to the Bank. From the on the communication between the Bank and it's client can be secure.

If this is so widely accepted, an alliance between the Public Sector and the Finance sector could leverage this existing infrastructure to distribute qualified, public certificates using the banks as Registration Authorities.

The current bank customers are not able to use their existing Bank certificates in communicating with other than the bank it self, it is a 1 to 1 relationship. In order to create a n-to-n relationship, the client needs a qualified certificate issued by a qualified Certification Authority. But it would be easy to apply for this certificate using the secure Home Banking communication.

This would immediately create millions of certificates in Europe, and would ease the work of many government organisations and at the time opening up for access to personal files and data stored in numerous databases.

It is to be expected, however, that next generation of PC's, PDA's and other devices will be equipped with devices that can read smart cards or similar devices, so that the portability aspects of a digital certificate could be ensured with much lower costs in just a year or two.

When the national Governments engage and encourage widespread use of digital certificates, the next bottleneck will most likely be the need for a practical European cross certification authority. It is highly recommended to solve this problem soon.

4.6 Common solution requirements

An e-records management solution should at a minimum have the following attributes:

- Ability to manage retention and disposition of records within an existing application
- No need for an additional repository for records
- Ensure records are destroyed when retention requirements are satisfied
- Non-intrusive to existing IT infrastructure;
- Non-intrusive to workers
- Ability to process as many as 1 million records per day in a single application

4.7 Emerging issues

Admissibility of e-records as evidence is dependent upon accuracy, authenticity, completeness, and trustworthiness. Audit Trails need to include information about who used the system, when did they use it, what did they do, and what was the result of their actions. Also, electronic signatures, recording and capturing as e-records content, context and structure are key issues.

Information and Records Managers must concern themselves with whether the content is correct and actually useful, what is the record is context & structure required to complete the record.

Traditionally, web-based information has fallen "outside the law" – not included in e-records retention practices for corporations. That needs to be dealt with in the future.

5. Standards for User Access and Information Protection

There are numerous standards available from the records management level and down to the level of system security and internal representation of data.

Also, the area of Digital Rights Management and Digital Media Management is very relevant for ensuring secure and reliable information dissemination and interchange covering any complexity of information from data records to video.

5.1 Model Requirements for the Management of Electronic Records (European Union)

The need for standard in the area of electronic records in the European Union was first articulated by the DLM-Forum, which has published the Model Requirements for the Management of Electronic Records (MoReq) in INSAR (Information Summary on Archives, Supplement VI) and on the following websites:

- http://europa.eu.int/ISPO/ida/
- http://dlmforum.eu.org
- http://www.cornwell.co.uk/moreq.html

Its scope is governmental as well as private organisations.

The requirements focus on function but address related issues like management of physical records as well. It can be used in its entirety or as components to a specification in a governmental organisation or a company and is thus relevant for any governmental organisation planning for records management.

5.2 Design Criteria Standard for Electronic Records Management Software Applications (USA)

This standard of the United States, Department of Defense (DoD) 5015.2 standard, describes the minimum functional requirements for electronic records management software applications to meet certain governmental regulations in the US. As a governmental body with a high degree of internal standardisation this standard can be relevant as inspiration and also because parts of it is being re-used by other nations.

5.3 Functional Requirements for Electronic Records Management Systems (United Kingdom)

The Public Record Office (PRO) of the United Kingdom is the national archive of England, Wales and the United Kingdom. Its mission is to preserve the records of central government and the courts of law from the 11th century until today.

Related to the e-government initiatives in the UK, the aim is be able to store, manage and retrieve their public records electronically by 2004.

The Public Records Office monitors the achievement of this target and acts as an advisory body to help governmental departments and agencies in developing their own specifications for electronic document and records management.

The specification is not a standard but rather a specification setting out the minimum requirements for electronic records management.

5.4 Functional Description, Requirements and Specifications for Record keeping Systems (Norway)

Noark-4 of the Norwegian National Archives is the fourth version of a specification of requirements for electronic archiving in governmental organisations. It describes requirements for

- Content (what kinds of information need to be recorded)
- Data structures (specification of data elements and their relationships)
- Function of a system

Although there are also some requirements regarding user interface this is by and large left to be decided by the specific solution provider. There are no specifications regarding implementation or systems architecture either.

The first version of Noark was introduced in 1984. It carries the specifications of Koark, a standard for local government practice.

Noark-4 describes a complete electronic archiving system that can be integrated with e-mail and electronic case management. Due to its breadth in scope it describes several tiers of implementation, allowing the organisation.

5.5 ISO 15489: Archives and Records Management

This international standard of ISO International Standard Organisation (http://www.iso.org) provides a high level framework for record keeping, including records management, regulatory considerations, and assignment of responsibilities. It covers capture, retention, storage, access, audit operations and staff training requirements.

There is a detailed guidance for implementing the records management framework, including policy and responsibility statements, management of the records management process, and various tools like security schemes and the use of thesauri.

5.6 ISO 5964: Establishment and Development of Multilingual Thesauri

This standard covers the use of multilingual thesauri to facilitate information interchange across national languages.

It builds upon the ISO 2788 standard for monolingual (i.e. single language) thesauri addressing the specific issues that are introduced dealing with more than one language all being considered as being equally important in terms of information provision.

With the EU being a multilingual environment the issue of structured information interchange across languages is important to EU member states.

5.7 ISO 11179: Specification and Standardisation of Data Elements

This standard specifies various aspects of data elements, including metadata. Metadata is data about data, i.e. descriptions of the various data elements in a system and how they relate to each other.

It defines a data element as consisting of

- Object class: description of a set of objects having similar properties
- Property: common peculiarities for all members of an object class,
- Representation: how data are represented, e.g. data type (numeric vs. text)

The standard also describes how classification schemes relate to data element concepts, and which attributes data elements have. Currently this model is being expanded to comply with other standards for metadata modelling
There are recommendations as to how to develop metadata definitions, and how these definitions can be registered with an external registration authority to facilitate information interchange across organisations.

This is a standard at a relatively high abstraction level addressing architectural issues closer to the operating system than many other standards. It can be used as guidance when designing information structures that are interchanged across systems.

5.8 LDAP - Lightweight Directory Access Protocol

Lightweight Directory Access Protocol (LDAP) is an Internet protocol for accessing directories. LDAP is an open industry standard that has evolved to meet the needs for accessing and updating information in directories.

In 1988, the CCITT (Consultative Committee on International Telephony and Telegraphy), created the X.500 standard, which became ISO 9594, Data Communications Network Directory, Recommendations X.500-X.521 in 1990, though it is still commonly

referred to as X.500. X.500 organises directory entries in a hierarchical name space capable of supporting large amounts of information.

LDAP requires the lighter weight and more popular TCP/IP protocol stack rather than the OSI protocol stack. LDAP also simplifies some X.500 operations and omits some of its less used features.

A directory is a listing of information about objects arranged in some order that gives details about each object. Common examples are a city telephone directory and a library card catalogue. In computer terms, a directory is a specialised database, also called a data repository, that stores typed and ordered information about objects.

Directories allow users or applications to find resources that have the characteristics needed for a particular task. For example, a directory of users can be used to look up a person's e-mail address or fax number. Searching a directory is similar to looking up a name in the white or yellow pages of a telephone directory.

However, directories stored on a computer are much more flexible than the yellow pages of a telephone directory because they can usually be searched by specific criteria, not just by a predefined set of categories. LDAP directories can be used for authenticating a user to network services.

LDAP is particularly relevant in a multi application and multi system environment where cross-system information protection is needed.

5.9 Security standards

Several standard bodies produce standards for security - including:

- Internet Engineering Task Force (IETF)
- American National Standards Institute (ANSI)
- Institute of Electrical and Electronic Engineers (IEEE) and
- International Standards Organisation (ISO).

Main industry consortia producing standards and guidances are:

- The Open Group (TOG)
- The Object Management Group (OMG),and
- World Wide Web Consortium (W3C)

DES - Data Encryption Standard and SET - Secure Electronic Transaction (TM) are examples of standards emerging from these bodies.

ECMA - the European Computer Manufacturers Association - does work in Standardising Information and Communications Systems. ECMA feeds draft standards into ISO.

IETF Security has Working Groups on e.g.

- Authentication, Authorisation & Accounting
- Common Authentication Technology
- Intrusion Detection
- IPSec
- DNS Security
- Authenticated Firewall Traversal
- PGP
- Transport layer Security
- Public Key Infrastructure
- Secure Shell
- S/MIME
- Web Transaction Security and
- Site Security Handbook.

5.10 Digital Rights Management and Digital Media Management - Standards, standardisation activities, fora and consortia

Despite the number of entries in this list it is in fact non-exhaustive. Digital Rights Management and Digital Media Management are emerging areas. They will have significant impact on information interchange in the future so a comparatively extensive coverage has been chosen in this chapter. Depending on specific project interests different sets of these standards will be relevant.

- IEC ITA OPIMA Specification 1.1
  The Open Platform Initiative for Multimedia Access (OPIMA) provides a standardised framework allowing the secure downloading, installation and running of proprietary protection systems (called IPMP systems).
- DVB-MHP (Multimedia Home Platform)
  DVB-MHP (ETSI TS 101-812) is a series of measures designed to promote the harmonised transition from analogue TV to a digital interactive multimedia future.
- ISO/IEC 21000 (MPEG-21)
  The scope of MPEG-21 could be described as the integration of the critical technologies enabling transparent and augmented use of multimedia resources across a wide range of networks and devices to support functions such as: content creation, content production, content distribution, content consumption and usage, content packaging, intellectual property management and protection, content identification and description, financial management, user privacy, terminals and network resource abstraction, content representation and event reporting.
- ISO/IEC 13818-1:2000 (MPEG-2) The MPEG-2 IPMP is designed so that they can be applied to any MPEG-2 multimedia representation. This is achieved by: (1) describing the functions that the IPMP Tools may perform; and (2) providing specific implementation of these functions on a particular distribution environment

- ITU-T SG 16 Question G
  ITU-T SG 16 Question G is working on multimedia security issues.
- MoU between ISO, IEC, ITU and UN/ECE on electronic business
  These organisations have signed an MoU in the field of electronic business to co-operate on standardisation of components involved in the areas of business scenarios, message and interoperability standards for business transactions, and product definition data standards for design, manufacturing and product support.
- E-SIGN Electronic Signatures Workshop
  While the EU Directive on Electronic Signatures provides a comprehensive legal framework for the harmonisation of the security infrastructures and authentication services using electronic signatures across Europe, it needs the appropriate technical support that will achieve the full implementation of its legislative prescriptions into the member states laws.
- W3C World Wide Web Consortium
  W3C is the organisation that develops primary specifications for the web. In particular, one of its major initiatives, XML, will form the basis for information interchange for the next generation of computer systems.
- PRISM Publishing Requirements for Industry Standard Metadata
  PRISM is an extensible XML metadata standard for syndicating, aggregating, post-processing and multi-purposing content from magazines, news, catalogues, books and journals.
- EbXML Electronic business using extensible markup language
  EbXML is a modular suite of specifications that enables enterprises of any size and in any geographical location to conduct business over the Internet.
- IDF International DOI Foundation
  The Foundation was created in 1998 and supports the needs of the intellectual property community in the digital environment, by the development and promotion of the Digital Object Identifier system as a common infrastructure for content management.
- OASIS Organisation for the Advancement of Structured Information Standards
  OASIS, creates interoperable industry specifications based on public standards such as XML and SGML, as well as others that are related to structured information processing.
- IRTF Internet Research Task Force
  IDRM Internet Digital Rights management is an IRTF Research Group formed to research issue and technologies relating to Digital Rights Management (DRM) on the Internet.
- OeBF OpenEBook Forum
  The Open eBook Forum is an association of hardware and software companies, publishers, authors and users of electronic books and related organisations whose goals are to establish common specifications for electronic book systems, applications and products.

- cIDf Content ID Forum
  cIDf is a forum to standardise "Content ID", which is a set of meta-data including a unique code embedded in each digital content item, providing content uniqueness and stabilising content value.
- TV Anytime Forum
  The TV Anytime Forum is an association of organisations that seeks to develop specifications to enable audio-visual and other services based on mass-market high volume digital storage.
- Schemas Forum
  SCHEMAS provides a forum for metadata schema designers involved in projects under the IST Programme and national initiatives in Europe.
- OeBPS OpenEBook Publication Structure
  The OeBF Format, developed to version 1.0.1, is increasingly used by publishers as an underlying file format for the production of derived proprietary eBook formats, such as Microsoft .lit and MobiPocket.

6. Best Practice Applications

6.1 Open Digital Administration Project of the cities of Naestved and Skurup

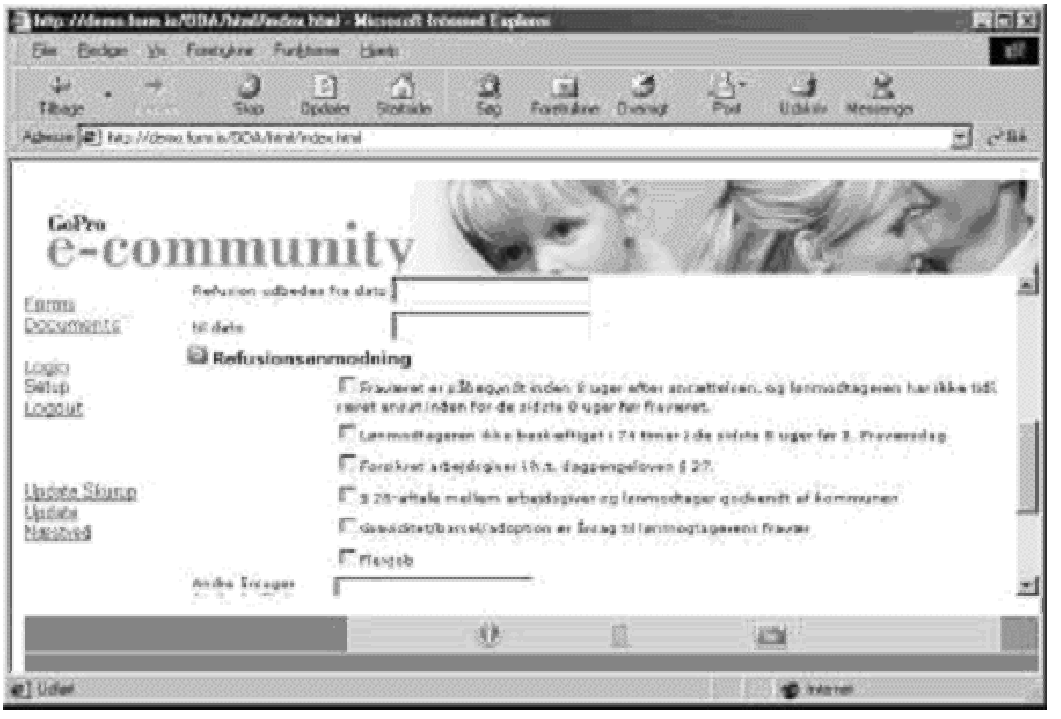


Figure 6: See http://www.naeskom.dk/webdatabaser/oda.nsf

An excellent illustration of the impact of improving user access to public administration can be found in the ODA project – Open Digital Administration. This project was started as a preparatory action project under the eContent programme in January 2001.

6.1.1 User organisation

The ODA consortium a joint project between the City of Naestved, Denmark and the City of Skurup, Sweden As technology providers the Icelandic Company Hugvit participates together with IBM. PBS – The Common Bank Service centre in Denmark – is issuing digital certificates on behalf of the cities and NALAD – the National Association of Local Authorities in Denmark participates with intelligent forms.

The cities Naestved and Skurup are where the actual project is implemented and operational. Naestved is a city of approximately 45.000 inhabitants, Skurup somewhat smaller, around 20.000 inhabitants.

6.1.2 Problem

The problem that the project is aiming to resolve is the lack of applications and experiences in secure, advanced and yet easy-to-use self service facilities for the citizens and the companies in dealing with the public sector.

More than 3.000 different forms are used in a traditional administrative system as the format for the citizens and companies way of interacting with the public. And when you investigate the forms, a lot of the data in the forms that to day has to be filled in by the citizens and companies consist of data already residing somewhere in the files of the public sector.

This creates a lot of additional work; it introduces a number of errors, reduces the turn around time and is generally regarded as poor public service.
The main objective of ODA is to demonstrate a robust, secure and scalable solution to offer self-service access to Public information.

As a test case in Denmark the project pilots the Danish act on Illness Insurance according to which an employer is refunded (part) of an employees salary if he/she has been ill for more than a certain period. In the city of Skurup the pilot project concerns building permits.

6.1.3 The Technical Solution

Backed by co-funding from the commission, the Consortium decided on the following major building blocks:

- Use of a Public key Infrastructure based on digital certificates
- Use of IBM Websphere Application Server as a web portal
- Intelligent forms provided by NALAD according to the Danish legislation
- The SERVEX workflow engine provided by Hugvit (developed under the ESPRIT IV programme EP 25639)
- The Lotus Domino back end server and the GoPro case management solution from Hugvit.

The concept of the ODA project is based on digital certificates, issued in accordance with the European Directive and the Danish Law on Digital certificates, citizens and companies can apply for refund by using intelligent electronic forms which can be handled by the Cities' document management and workflow systems and automatically transfer to the mainframe data systems for processing.

The City of Naestved has the formal role of Certificate Authority and based on the key software ingredients that are accepted by the Danish National Procurement Agency as part of a tender for PKI software solutions, the Certificates are issued by PBS using Tivoli PKI and the registration function takes place in Skurup and in Naestved based on the same

Software as the banks are using for Home banking, called CBT – Crypto Based Transactions from IBM.

Because of the use of digital certificates, the employer and the employee are now able to access a web site and fill in only the data needed for the actual case, not all the other information, which is already stored somewhere in the public system regarding the company and the citizen in question. The intelligent form checks for correctness and using an XML interface it is automatically forwarded to the Cities' workflow and journaling system and at the same time updating the back end databases.

The forms that have been piloted so far were developed using XML. When the end user signs on to the ODA website and present his digital certificate, the content of the form is dynamically collected from the files residing at the City server, and on completion the form is signed by the citizen, data is automatically forwarded to central databases and entered into the case management system located at the Notes server.

6.1.4 Experience

The project was officially put into operations at the end of October 2001 following a 9 months definition and development phase. The total effort for this project is 143 man months and the total cost estimated to 1.6 mio Euro.

The first two key applications put into operation was the reporting and claim for illness reimbursement from the state in Denmark and the administration of building permits in Skurup.

In both cases the effects have been significant. The type of case processing covered by the project in the city of Naestved, big data need to be entered into the traditional application form and this in itself creates a large number of errors. As much of 40% of the traditional forms contain errors leading to tedious and labour-intensive corrections and delays in the process.

The Naestved experience points towards a reduction in the error rate of the forms from more than 40% of the traditional forms being erroneous to now 0 errors. The case processing time has been reduced from weeks to a few days cutting the administrative costs by more than a third.

In the city of Skurup similar benefits can be noted in terms of a huge reduction in errors and an impressive reduction in response times. In both pilot projects the combination of improved service and reduced administrative cost are noteworthy.

At the same time the procedure of issuing and using digital certificates has proved to be without major problems and as the certificates are based on soft certificates, no installation problems has been noted for the end users.

The promise for this type of applications is impressive as a total number of 1300 forms are estimated to cover more than 80% of the number of transactions between the Public Sector and the citizens/companies

6.1.5 Adaptability

The general adaptability of the solution is that both cities are now ready to expand the experiences to the other more than 2.000 forms and in this way greatly reduce administration costs. These applications could of course be used by any other city in Denmark and Sweden, respectively.

Even if the forms are national by nature, the general infrastructure could be copied in any other European country – the Digital Certificates are issued in according to the Danish/Swedish laws, derivatives of the European Digital Signature Directive, and the administrative procedures for local government are not that different from country to country, although legal and language translation has to be performed.

6.2 Personal Portal Solution The Keen Project

6.2.1 User organisation

KEEN – Knowledge for Everybody Everywhere on the Net – was started in July 2001. Three cities in the Nordic countries - Naestved in Denmark, Skurup in Sweden and Arendal in Norway - joined together with technical partners with the intention of piloting applications and infrastructure based on the experiences of the ODA project but with a vision to take this even further.

6.2.2 Problem

The objective is to provide a flexible, scalable platform based on 4 visions:

- Content only exists on the NET
- Everybody should have access to services, knowledge, education, decision-support (tools) and entertainment from anywhere based on their own choice of channel be it telephone, television, personal computer, or wap-station.
- Intelligent “sniffer-tools” is an end user development tool designed to assist the individual in managing preferences and choices. It starts from a local well-defined and prioritised starting point like a city Home Page Portal offering a limited set of secure services in a preferred language. It then extends to a new, dynamic, individually organised and defined personal portal that will enable the individual with the power to access and benefit from all services and information possibly provided on the net by other providers - be it other local authorities, states, organisations, business entities. In this way it vastly expands the reach and range of local information content and services available in the physical neighbourhood.

- By combining digital signature with payment systems and IPR-protection systems provide a generally accepted and practical tool to attract every relevant participant in the mindscape.

6.2.3 Technical Solution

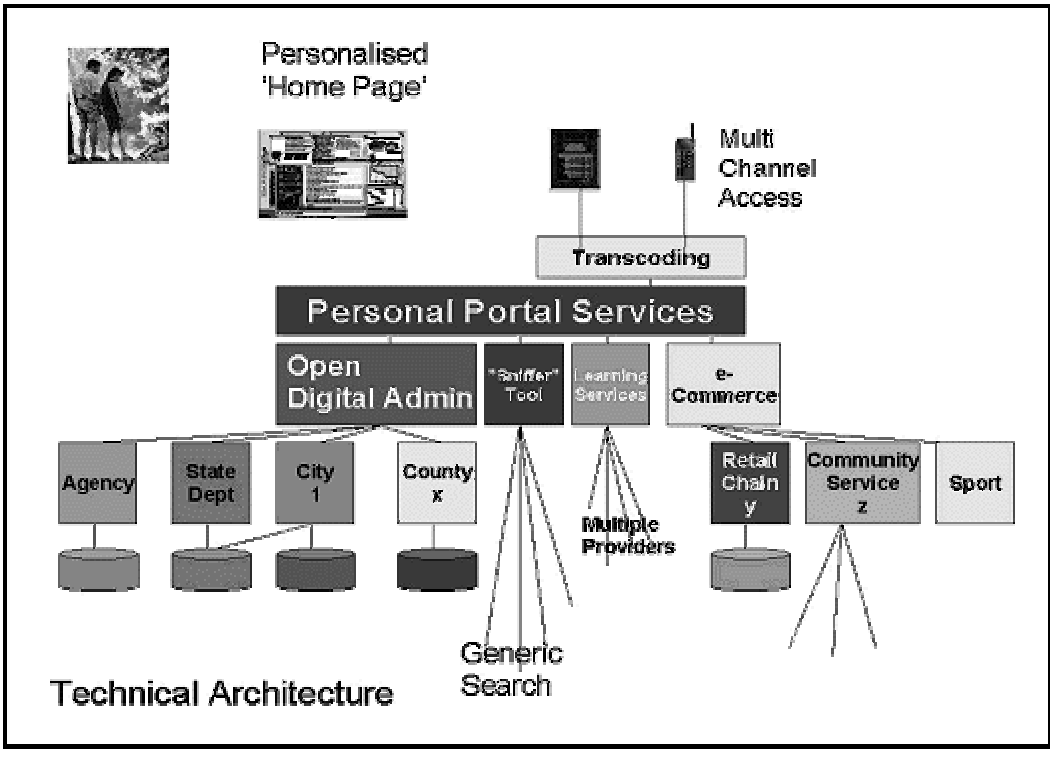


Figure 7: Technical architecture

The IT architecture as shown above consists of the following elements:

- A personal portal – or more precisely – a web homepage with dynamic, user defined content, a community portal connecting the citizen to a number of services and with the capability of converting and transcoding the information so that the citizen can use a variety of devices or channels,
- A sniffer-tool – a generic search tool to assist the citizen to find relevant information – a replacement for application development
- Learning Services – covering 2 different types of services: General learning spaces offering both tools and distance learning and Learning Village, a solution specifically aiming at bringing pupils, teachers and parents together

The concept also includes access to Government services, including knowledge-oriented services like information on legislation, citizens' rights, and support for dealing with public administration.

To make this a generic offering the solution can be expanded to include also commercial types of offerings like e-commerce, or leisure-oriented services supporting interest groups like sports clubs etc.

This infrastructure is backed by security systems involving digital certificates to ensure authenticity of the participating partners and privacy and confidentiality where needed
It is also the intention to include micro-payment services to solve the payment side of intellectual property rights of the content offered through this infrastructure.

In order to support citizens' use of multiple channels, the partners will also test the use of transcoding, and by applying Speech Synthesis, it will be possible for people without computers to use Mobile Phones to access information on the progress of their cases with the public sector – or to listen to information retrieved on the Internet.

6.2.4 Experiences

This project, with the working title KEEN – Knowledge for Everybody Everywhere on the Net – will start implementation in the spring of 2002 as a joint effort between the cities of Naestved, Denmark, Skurup in Sweden and Arendal, Norway.
What is expected to come out of the project is an evaluation of the way this will influence how citizens react, live and feel involved and in control of their own situation. Also it is hoped to demonstrate that a much easer access to training and education will help encourage social inclusion and improve employability.

6.2.5 Adaptability

The project will be divided into focused, well-defined subprojects for which each of the cities will take responsibility. The objective of course calls for 'cross fertilisation' and as the objective also is to describe the European perspective, each application developed and piloted in one city will at the end of the project be implemented and tested in the other two cities.

The way this project is established is in fact turning the Public Administration upside down; traditionally what is called 'digital administration' or even eGovernment looks at the administrative process and the dissemination of information and knowledge from the Public Sector viewpoint.

The KEEN project puts the citizen in the centre and behind the driver's wheel. You could describe this as the second Copernican revolution and there is no doubt that to many public sector traditionalists this will feel like a revolution.

6.3 Enterprise Content Management System Project of the Statens Museum for Kunst – The National Danish Art Museum

6.3.1 Description of the organisation

Statens Museum for Kunst/The National Danish Museum of Art, that is located in Copenhagen, is the Danish national art gallery. It's collections cover the Danish historical and contemporary collections of paintings, sculpture and art works. A major part of its collection comes from the art chambers of the Danish Kings. Its main building is an architectural art piece in itself, created by the Architect Wilhelm Dahlerup in 1896 with a modern extension built in 1998.

6.3.2 Problem

The collection counts more than 400.000 pieces and at any point in time only about 2.000 can be exhibited to the Public. The board of directors from the museum for many years have been observant to this problem and started some years ago the tedious task of describing the collection using XML-format and storing JPEG-pictures of the collection.
The challenge was to make this database accessible to researchers as well as normal citizens in Denmark or abroad in a way which made searching of the database meaningful and fast.

6.3.3 Technical solution

The solution chosen for this application is based on the IBM Content Manager software, including IBM Enterprise Information Portal Client Kit, IBM DB/2 Universal Database and Lotus Domino. The hardware solution includes a firewall and the server is running on an IBM Netfinity 5500.

The IBM Content Manager accepts the XML format for the textual description of the archives as well as the JPEG format for the digital pictures of the paintings. The Content manager program offers text search features that are intelligent in the sense that misspelled words or phrases are 'translated', making it ease for the casual user to search through the collection.

When a user submits a search form, the Lotus Domino Server sends the data using a Java servlet through the Enterprise Information Portal Client to the IBM Content Manager, who then passes the search form to it's index list of matching items. The Enterprise Information Portal includes a number of search technologies, content repository connectors and server-based transforms.
When an indexed item is selected, the content connectors provide a fast access to the stored text and picture.

The Content Manager assists the administrative personnel responsible for describing and archiving the data using DB/2 as an index database.
The main criteria for the choice of technical solution were the ease-of-use and the scalability and robustness of the solution. From the opening in the beginning of 2001, 5.000 items where stored in the database and as this number will expand in a few years to cover a major part of all 400.000 items, both the database and the potential number of users would require the possibility of upgrading the solution to a mainframe.

6.3.4 Experiences

The solution receives more than 50.000 hits per week, allowing visitors also to educate themselves, check the museums calendar and sign up for special events.
The management of the National Gallery has been very pleased with the solution, provided by IBM Global services and a business partner, Semaphor.

The staff's vision will be to provide even higher quality digital images to allow for zooming capabilities to study fine details, a 3D walking experience in the virtual Gallery using IBM HotMedia. To day, 6 months after opening the site is also linked to an on line bookstore selling books of art from the Gallery's bookshop.

6.3.5 Adaptability

Similar experiences and solutions can be applied to other types of museums – or similar art museums. Building on the experience from this case, from for instance the solution used at the Hermitage Museum in St. Petersburg and from the Vatican Archives, the use of advanced content management solutions may well change the way researchers are working with our cultural heritage and provide a lot of pleasure and education for the citizens of Europe.

7. Outlook

7.1 Proven strategies

Scalable, n-tier architecture based solutions focusing on ease-of-use for the citizen as well as for the government have proven successful.
Using Digital Certificates open up for the individuals securely to access and even change their own data and storing it in an encrypted form, only available to the case officer and themselves. The users can access a web site and fill in only the data needed for the actual case, not all the other information, vastly extended availability of information for the citizens. This has led to a significant reduction in error-rates and case processing times.

7.2 Technology benefits

With the rapid growth of information today, text-mining techniques are key for the ability to navigate and extract the information needed. Those techniques also assist in categorising and summarising the information.
Besides ensuring the integrity of the governmental processes, records management technology protects private information and counters cyber crime - support for pervasive devices.
The most profound change the usage of these technologies will imply is that the citizen will be put in the centre transforming many governmental processes from government-driven to citizen-driven. This will influence the way the citizens react, live and feel involved and in control of their own situation.

7.3 Critical Success Factors

User and organisational requirements take priority over opportunities to apply technology. An appropriate balance between access and information protection is needed. Provide maximum access to information to those needing and requiring it, while at the same time securing that same information from unauthorised access.
Manage the retention of information. Ensure you are retaining what you are required to retain and destroying what you can, when you can. This will help build credibility and integrity within systems.

7.4 Trends

Several technologies will drive future changes in user access and information protection. Key drivers are:

- from client-server to n-tier architectures
- Growth in information volumes
- Increased bandwidths
- Growth in pervasive, mobile devices
- Portal and information mining technologies, including voice technology

The world of IT is shifting very rapidly from traditional client-server technology to n-tier architecture, at this time and is expected to be pervasive with two years or less. All industry leaders are already there or have a strategy to be there in six months or less, e-government systems being no exception.

3-tier or n-tier architecture includes:

- An management tier including: actual content and data resources;
- An application software tier including: application services; and
- An information delivery tier: browser-based, that can be accessible either via desktop workstation or other components, such as a PDA.

The mere possibility of storing and providing information will drive the information volumes up even further. There is various research on the expected growth of information in the future. They all deal with the almost impossible task of quantifying information in a meaningful way, and we will refrain from quoting specific studies in this chapter. In general, these studies express the following view of the current situation and expected growth:

- Today's IT systems only leverage a minimal fraction of the information available. This is true for organisations as well as the world as such. (sample statement: only 0,5% of the information available in recorded form in the world is available online)
- Information volumes will grow explosively in the future (sample statement: in the future, the global information volumes will double every 3 years)

With the constant reduction in storage and network costs, it will become attractive to record and use even more information than we are doing today.
A larger part of the information volumes in the future will be digital media like high-resolution images, moving pictures and sound.

Speech technology combined with text mining and portals will make it possible to search for information delivered in meetings in an intelligent way. Obviously this will introduce new requirements for security, access control and protection of intellectual assets.

These trends will change our world profoundly. Think of what we have achieved managing the current tiny fraction of the world's information: ERP applications, government archives, databases, e-commerce solutions, Customer Relationship Management, e-mail systems, and the entire Internet. Extrapolating this to the benefits gained from leveraging information volumes many times larger goes beyond even the most visionary mind.

Glossary

ADL (Advanced Distributed Learning)	ADL is an initiative by the U.S. Department of Defence to achieve interoperability across computer and Internet-based learning courseware through the development of a common technical framework, which contains content in the form of reusable learning objects.
Associative Access	Knowledge retrieval based on pattern matching between an unstructured query (text paragraph) and a document content store.
Authoring tools	Tools/SW to create and adapt content to the web for use in an online course. They assist in creating e-learning solutions and provide a "do-it-yourself" option for placing content and materials online.
Categorization / Category	Assigning documents to different groups by performing content-related analysis - so called categories. Categorization schemes are typically built upon business processes and business rules or rely on knowledge domains within an organization.
CD-ROM assessment	An assessment or survey that can be accessed and completed by using a CD-ROM launched through a company's intranet. CD-ROM based assessments also can be used on a desktop stand-alone computer if the assessment is a self-assessment for the benefit of the trainee only. Alternatively, a CD-ROM-based survey can be printed (if the CD-ROM has a print capability) and used as a paper-based survey.
Computer-based training	A term used to describe any computer-delivered training, including CD-ROM, the Internet and Intranets. Sometimes referred to as Computer-assisted instruction (CAI), CBT is asynchronous learning.
Classification / Class	Collection of methods applied to categorize documents by analysing their content. In many cases, categories and classes are identical. Categories incorporate the semantics of the application, whereas classes may also be of formal nature.
Classify	Classification is a method of assigning retention/disposition rules to records. Similar to the Declare function, this can be a completely manual process or process-driven, depending on the particular implementation. As a minimum, the user can be presented with a list of allowable file codes from a drop-down list (manual classification). Ideally, the desktop process/application can automate classification by triggering a file code selection from a property or characteristic of the process/application.
Content Search	Information retrieval based on pattern matching between a query (text paragraph) and a document repository.
Distance learning/ Interactive Distance Learning (IDL)	Traditionally refers to a broadcast of a lecture to distant locations, usually through video presentations. IDL is a real-time learning session where people in different locations can communicate with each other. Videoconferencing, audio conferencing or any live computer conferencing (e.g., chat rooms) are all examples of IDL.
Document	A document (any form or format), an email message or attachment, a document created within a desktop application such as MS Word, regardless of format. There are two forms of document: Electronic Document: Body (text) of the document is stored in electronic format and can be read. If declared as a record, an electronic document becomes a managed record (i.e. a document may or may not be a (declared) record) Non-Electronic Document (Ndoc): A physical document of any form (maps, paper, VHS video tapes, etc.). Body is not recorded in electronic form, but descriptive metadata is stored and tracked within CM (profile). If declared as a record, an Ndoc becomes a managed record (i.e. an Ndoc may or may not be a (declared) record).
Document Life Cycle Management	The records life cycle is the life span of a record from its creation or receipt to its final disposition. It is usually described in three stages: creation, maintenance and use, and final disposition. e-Records applies management to all three stages. With e-Records, the records manager can create and maintain the official rules that will dictate when to destroy (or permanently keep) electronic records, as well as record

	and enforce any conditions that apply to destruction (e.g. destroy 2 years following contract completion). Finally, the records manager can carry out the physical destruction of electronic records, maintaining a legal audit file.
Document Security Control	Access control to documents (non-declared records) Note: Document security control is different from Records Security Control.
Electronic Recordkeeping	The practice of applying formal corporate recordkeeping practices and methods to electronic documents (records).
Electronic Signature	A signature is a bit string that indicates whether or not certain terms occur in a document.
Enterprise Content Management	Manage all content (i.e. unstructured information) relevant to the organisation. It embraces three historically separate technologies: web content management, document management, and digital media asset management. While outwardly dissimilar, all of these forms of enterprise content share similar needs for mass storage, search and access, personalisation, integration with legacy applications, access and version control, and rapid delivery over the internet.
EPSS (electronic program support system)	An electronic system that provides integrated, on-demand access to information, advice, learning experiences and tools. In essence, the computer is providing coaching support (i.e. the principal of technology based knowledge management).
File	A disk “file”, something stored on electronic media, of any file. Does not necessarily denote a record. For example, “image files are stored an a server” simply refers to the electronic images, and implies nothing about the records status. Will be used in the context of describing the storage of documents and related information to electronic media.
File Plan Administration	Design and administration of the corporate file plan. The records manager can design file plan components. With Tarian’s file plan designer, the records manager can design classes of file plan objects (files, records, folders, etc), then define the attributes of these classes. Relationships between classes are then defined (i.e. files can contain files, records and folders). Various views of the file plan may be defined. For instance, a warehouse view might present a view of the physical folders in the organization, whereas a numeric view might present the sorted numeric structure for maintenance purposes. The records manager can create pick-lists enforcing consistency within the file plan, component profiles that define the characteristics of the file plan, and default values to simplify daily file creation tasks. Policies, Permissions, and Suspensions may be assigned to file plan objects.
Information mining	Linguistic services to find hidden information in text documents on content servers
Information Retrieval	An information retrieval (IR) system informs on the existence (or non-existence) and origins of documents relating to the user’s query. It does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. This specifically excludes Question.
Keyword Search	Information retrieval method based on literal match of words.
Learning Resource interchange (LRN)	LRN is the Microsoft implementation of the IMS Content Packaging Specification. It consists of an XML-based schema and an LRN toolkit. It enables a standard method of description of content, making it easier to create, reuse and customise content objects with an XML editor, whether initially developed from scratch or bought under license from vendors.
Neural Networks	In information technology, a neural network is a system of programs and data structures that approximates the operation of the human brain. Typically, a neural network is initially "trained" or fed large amounts of data. A program can then tell the network how to behave in response to an external stimulus (for example, to classify a document based on its content).
Pattern Matching/ Recognition	Matching/Recognition of objects based on features. Pattern Matching with regard to text documents means to identify and match words and phrases from different documents under the assumption that the more features match, the more similar the contents are.
Personalisation	The ability to provide the user with the right content both from the user's and Web

	site owner's perspective. A personalization algorithm determines whether content is presented to the user, and if so, in what order of priority.
Portal	A single integrated point of comprehensive, ubiquitous, and useful access to information (data), applications, and people.
Record	Any form of recorded information that is under records management control. Records are either Physical or Electronic. Records may take any of the following four forms: Document: A document (see above) that has been declared as a record. Once declared as a record, the document is under records management control Folder: A folder of (paper) documents. Individual documents within the folder may or may not be treated as records (declared Ndocs). The physical handling of the folder is managed by Tarian's Physical Records Module Box: A box of (typically) paper documents. Usually contains folders (see above), which are individually managed as records, but may alternatively contain records other than folders such as loose documents of a given subject. The physical handling of the box is managed by Tarian's Physical Records Module Ndoc: A declared Ndoc (See above for definition of Ndoc) Important: A document (electronic or Ndoc) will not be considered to be a record until has been declared.
Record, Electronic	Electronic Records (e-Records). Any information (document) recorded in electronic form, on any digital media, that has been Declared to be a record. Characteristics of an e-Record: Document is in electronic form Metadata is associated with the document Document has been classified against a file plan Only the authorised Records manager has the means by which to apply retention/disposition to the document.
Record, Physical	Folders, Boxes, Ndocs to which records management control has been applied. A document (electronic or Ndoc) becomes an e-Record only once it has been declared.
Records Administration	The administrative infrastructure represents the tasks that the records manager carries out on the entire organization's collection of declared records. Conducted within Tarian's Records Administration Client, a browser-based web application. End users never see this process. Consists of the following four broad activities; File Plan Administration, Records Security Control, LifeCycle Management, and Reporting.
Records Manager	Conducts one or more records administrative functions.
Records Security Control	Access control to declared records. Users and Groups of users may be created, and assigned roles and policies that will interact to determine the records users are able to access. Note: Records security control is different from Document Security Control.
Reporting	The process of generating reports from data managed by eRecords solution. It is a tow-step process. Reports are first designed, and the design is saved for later reuse. Second, reports are generated by running the report design against the data.
Repository	Physical storage are for documents and/or electronic records.
Retention Rules	(Retention Schedule). The set of rules which specify how long to keep (retention) records, and what to do with them at the end of their lifecycle (disposition).
Syntactical Analysis	Syntactical analysis derives the syntactic category of words or phrases based on (language dependent) dictionaries and grammars. Example: house – noun.
Thesaurus	A book that lists words in groups of synonyms and related concepts.
Volume	Folder. A Volume will be referred to as a folder (common US terminology).
Virtual Reality (VR)	Virtual Reality simulations (usually involving wearing headgear and electronic gloves) that immerse users in a simulated reality that gives the sensation of being in a three-dimensional world.

Abbreviations

ASP	Application Service Provider
AVI	Audio Video Interleaving
BCR	Bar Coding
BPM	Business Process Management
CBT	Computer Based Training
CCD	Charge Couple Devices
CM	Content Management
COLD	Computer Output to Laser Disk
COM	Component Object Model
COOL	Computer Output On Line
DBMS	Database Management System
DMS	Document Management System
DRT	Document Related Technologies
ECM	Enterprise Content Management
E-Learning	Education, training and structured information delivered electronically
ERM	Enterprise Report Management
ERP	Enterprise Resource Planning
E-Term	European programme for Training in Electronic Records Management
FDDI	Fibre Distributed Data Interface
GIF	Graphic Interchange Format
HTML	Hypertext Mark-up Language
ICR	Intelligent Character Recognition
ICT	Information and Communication Technology
IDM	Integrated Document Management
ISDN	Integrated Services Digital Network
ISO	International Standards Organisation
JPEG	Joint Photographic Experts Group
KM	Knowledge Management
LAN	Local Area Network
LDAP	Lightweight Directory Access Protocol
MoReq	Model Requirements for the management of electronic records
MPEG	Moving Pictures Expert Group
NAS	Network Attached Storage
OCR	Optical Character Recognition
ODCB	Open Database Connectivity
OLE	Object Linking & Embedding
OMR	Optical Mark Recognition
PDF	Portable Document Format
PPP	Point-to-Point Protocol
RMS	Records Management System
RTF	Rich Text Format
SAN	Storage Area Networks
SQL	Structured Query Language
TCP/IP	Transmission Control Protocol/Internet Protocol
TIFF	Tag Image File Format
WAN	Wide Area Network
WAV	Audio Format File
WCM	Web Content Management
WebDAV	Web-based Distributed Authoring & Versioning
WORM	Right once read many times
XML	eXtensible Mark-up Language

References

MoReq INSAR	http://dlmforum.eu.org http://www.cornwell.co.uk
United States, Department of Defense (DoD) 5015.2 Standard	
United Kingdom, Public Records Office (PRO)	http://www.pro.gov.uk/recordsmanagement/eros/invest/default.htm
National Archives Norway, NOARK-4	http://www.riksarkivet.no/Noark-4/Om-Noark.htm
ISO 15489	http://www.iso.org
ISO 5964	http://www.iso.org
ISO 11179	http://www.iso.org)
IEC ITA OPIMA Specification 1.1	http://leonardo.telecomitalialab.com/opima
DVB-MHP	http://www.mhp.org
ISO/IEC 21000 (MPEG-21)	http://mpeg.telecomitalialab.com/standards/mpeg-21/mpeg-21.htm
ISO/IEC 13818-1:2000 (MPEG-2)	http://mpeg.telecomitalialab.com/standards/mpeg-2/mpeg-2.htm
ITU-T SG 16 Question G	http://www.itu.int/ITU-T/studygroups/com16/questions.html
MoU between ISO, IEC, ITU and UN/ECE on electronic business	http://www.unece.org/press/00trad1e.htm
Electronic Signatures (E-SIGN) Workshop	http://www.cenorm.be/isss/Workshop/e-sign/Default.htm
World Wide Web Consortium (W3C)	http://www.w3c.org
PRISM	http://www.prismstandard.org

EbXML Foundation	http://www.ebxml.org
International DOI(IDF)	http://www.doi.org
OASIS	http://www.oasis-open.org
IRTF Internet DRM WG	http://www.idrm.org
OpenEBook Forum	http://www.openebook.org
cIDf (Content ID Forum)	http://www.cidf.org
TV Anytime Forum	http://www.tv-anytime.org
Schemas Forum	http://www.schemas-forum.org
OeBPS	http://www.openebook.org/oebps/oebps1.0.1/download

Authoring Company

IBM

At IBM, we strive to lead in the creation, development and manufacture of the industry's most advanced information technologies, including computer systems, software, networking systems, storage devices and microelectronics. We translate these advanced technologies into value for our customers through our professional solutions and services businesses worldwide.

Contact

International Business Machines Corporation
New Orchard Road
Armonk, NY 10504 - USA
Tel: +1 (0)914 499 1900

IBM Eurocoordination
Tour Descartes
2, Avenue Gambetta
F-92066 Paris La Defense - France
Tel: +33 (1)4188 6000

Contact Authoring Company

IBM
Kim Jasper
Nymoellevej 91
DK-2800 Lyngby - Denmark
Tel. +45 (0)4586 5471
E-Mail: JASPER@dk.ibm.com

Industry White Paper Series on Records, Document and Enterprise Content Management for the Public Sector

Capture, Indexing & Auto-Categorization
Intelligent methods for the acquisition and retrieval of information stored in digital archives
ISBN 3-936534-01-2

Hewlett-Packard GmbH
Conversion & Document Formats
Backfile conversion and format issues for information stored in digital archives
ISBN 3-936534-02-0

FileNET Corporation
Content Management
Managing the Lifecycle of Information
ISBN 3-936534-03-9

IBM
Access & Protection
Managing Open Access & Information Protection
ISBN 3-936534-04-7

Kodak
Availability & Preservation
Long-term Availability & Preservation of digital information
ISBN 3-936534-05-5

TRW Systems Europe / UCL - University College London / comunicando spa
Education, Training & Operation
From the Traditional Archivist to the Information Manager
ISBN 3-936534-07-1

Publishing Information

The series of six Industry White Papers is published to address the needs of public administration and archives at the European, national, federal and local level and to educate the public sector throughout Europe about available solutions for archival problems on relevant topics about acquisition, management, long term storage, multilingual access, indexing and training issues.

DLM-Forum
The current DLM acronym stands for *Données Lisibles par Machine* (Machine Readable Data). It is proposed that after the DLM-Forum 2002 in Barcelona this definition be broadened to embrace the complete "**D**ocument **L**ifecycle **M**anagement". The DLM-Forum is based on the conclusions of the Council of the European Union, concerning greater co-operation in the field of archives (17 June 1994). The DLM-Forum 2002 in Barcelona will be the third multidisciplinary European DLM-Forum on electronic records to be organised. It will build on the challenge that the second DLM-Forum in 1999 issued to the ICT (Information, Communications & Technology) industry to identify and provide practical solutions for electronic document and records management The task of safeguarding and ensuring the continued accessibility of the European archival heritage in the context of the Information Society is the primary concern of the DLM-Forum on Electronic Records. The DLM-Forum asks industry to actively participate in the multidisciplinary effort aimed at safeguarding and rendering accessible archives as the memory of the Information Society and to improve and develop products to this end in collaboration with the users.

European Commission SG.B.3
Office JECL 3/36, Rue de la Loi 200, B-1049 Brussels, Belgium
A/e: dlm-forum@cec.eu.int

AIIM International - The Enterprise Content Management Association
AIIM International is the leading global industry association that connects the communities of users and suppliers of Enterprise Content Management. A neutral and unbiased source of information, AIIM International produces educational, solution-oriented events and conferences, provides up-to-the-minute industry information through publications and its industry web portal, and is an ANSI/ISO-accredited standards developer.
AIIM Europe is member of the DLM-Monitoring Committee and co-ordinates the activities of the DLM/ICT-Working Group.

AIIM International, Europe
Chappell House, The Green, Datchet, Berkshire SL3 9EH, UK
http://www.aiim.org

Industry White Paper Series on Records, Document and Enterprise Content Management for the Public Sector

The Industry White Papers are published by the DLM-Forum of the European Commission and AIIM International Europe to address the needs of public administration and archives at the European, national, federal and local level and to educate the public sector throughout Europe about available solutions for archival problems on relevant topics about acquisition, management, long term storage, multilingual access, indexing and training issues. The leading suppliers of Enterprise Content Management technologies participate in this series and focus on electronic archival, document management and records management for the public sector in the European Community.

Access & Protection

In this White Paper the key topics for user and information access will be addressed. Issues regarding litigation, privacy protection and networks attacks need to be addressed in order to provide secure access to citizens. The ability to locate and identify relevant information is becoming key - with the portal as a paradigm for the rich function needed for information access. Which standards are relevant to user access and information access? Planning for any significant IT application requires knowledge about standards – in particular with open application that will interact with many other systems. Protection of public information is not only about how to avoid hacker attacks. Governments need validated audit trails of their information interchange with their citizens, and there is a need for building proof of authenticity into the information infrastructure. The paper will also describe the main drivers for architectural change.

ISBN 3-936534-00-4 (Series)
ISBN 3-936534-04-7